Expert Systems with Applications 39 (2012) 10153–10159

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

CART-based selection of bankruptcy predictors for the logit model

Arjana Brezigar-Masten [a,b], Igor Masten [c,*]

[a] University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technology, Glagoljaška 8, SI-6000 Koper, Slovenia
[b] Institute of Macroeconomic Analysis and Development, Gregorčičeva 27, SI-1000 Ljubljana, Slovenia
[c] University of Ljubljana, Faculty of Economics, Kardeljeva pl. 17, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Bankruptcy prediction
Model selection
CART

ABSTRACT

Balance-sheet data offer a potentially large number of candidate predictors of corporate financial failure. In this paper we provide a novel predictor selection procedure based on non-parametric regression and classification tree method (CART) and test its performance within a standard logit model. We show that a simple logit model with dummy variables created in accordance with the nodes of estimated classification tree outperforms both standard logit model with step-wise-selected financial ratios, and CART itself. On a population of Slovenian companies our method achieves remarkable rates of precision in out-of-sample bankruptcy prediction. Our selection method thus represents an efficient way of introducing non-linear effects of predictor variables on the default probability in standard single-index models like logit. These findings are robust to choice-based sampling of estimation samples.

1. Introduction

The problem of predicting corporate failure is at the hart of risk management procedures in banks worldwide. The turmoil in financial markets after the outbreak of the financial crisis in 2008 has only re-emphasized the importance of proper risk assessment. Characteristics of a good bankruptcy prediction model would in this respect be: reliability and robustness, ease of implementation, high degree of prediction accuracy and clear interpretability and transparency of decision making process. The aim of this paper is to propose a bankruptcy prediction methodology that fulfills all of these characteristics.

Methodological approaches used in bankruptcy prediction can be broadly classified into statistical methods and artificial intelligence methods (Min & Jeong, 2009). The first group includes discriminant analysis (used in pioneering studies of Beaver (1966) and Altman (1968)) and binary-choice models like logit or probit. A common statistical property of these methods is that they are fully parametric. The artificial intelligence group comprises of methods that range from artificial neural networks (ANN) and genetic algorithms (GE) to classification and regression trees (CART) (see Li, Sun, & Wu, 2010; Paliwal & Kumar, 2009). A common statistical feature of this second class of models is a fully non-parametric specification of both the distributional form of variables and functional relations among them.

This paper attempts to bridge the two model classes and proposes a method of selecting predictor variables for the classic logit model based on the non-parametric classification and regression tree method.[1] The spirit of our approach is the following. We apply CART to a large set of possible predictors to estimate a decision tree that partitions firms into bankrupt and healthy. In accordance with decision nodes given by the tree we construct a set of dummy variables that are used as predictors in the logit models. Such CART-determined dummy variables can enter the logit model both independently or as additional explanatory variables to a set of conventionally selected variables. Our dummy variable approach is different from the approach of Cho, Hong, and Ha (2010) who used the (untransformed) variables that CART method reports in the estimated tree as input variables in several models, including the logit model. The use of dummy variables is both simple and efficient because it preserves the nonlinearity in the relations among variables identified by CART also in the single-index models. It is especially the potential non-linearity in the effect of candidate predictor variables on the probability of financial distress that conventional selection methods or the approach of Cho et al. (2010) cannot capture.

Our results demonstrate that our CART-based selection procedure of bankruptcy predictors is indeed a very useful method for selection of bankruptcy predictors, which combined with the standard logit model provides an accurate prediction tool. Validity and

* Corresponding author.
E-mail addresses: arjanabm@gmail.com (A. Brezigar-Masten), igor.masten@ef.uni-lj.si (I. Masten).

[1] Corporate balance sheets and income statements, which are the main data source for such application, can be used to compute numerous financial ratios, all of them in principle being candidate predictors of financial distress.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.125

robustness of this finding is corroborated by the fact that our comparisons of prediction accuracy are performed truly out of sample. In addition, we do not look only at overall prediction accuracy but put special emphasis to prediction accuracy of bankrupt and healthy firms separately. Moreover, our empirical findings are obtained on a population of Slovenian enterprises, ranging from very small private businesses to large international corporations. The final test of robustness of our findings is against the construction of the estimation sample. Namely, choice based sampling of the observations into estimation sample that equates the number of bankrupt and healthy firms in the estimation sample is a common approach in the literature and practice. While such an approach may be motivated by data availability and econometric considerations, it is definitely at odds with composition of data in reality and, consequently, bankruptcy prediction in practice. For this reason we test our method not only on a matched sample, but also on a large sample with population shares of bankrupt firms, which is the situation that banks face in real life.

The remainder of the paper is organized as follows. Section 2 introduces competing bankruptcy prediction models. Section 3 outlines our selection method and other methods with compare it to. Section 4 presents our data and the construction of estimation samples. Section 5 contains the results of predictor selection, while Section 6 reports the results of estimated logit models. Comparison of bankruptcy prediction is given Section 7. Finally, Section 8 concludes.

2. Prediction models

Our basic model of the probability of bankruptcy is the logistic regression. The logit model has been extensively applied in the literature (see for example, Chen, 2011; Li et al., 2010; Min & Jeong, 2009, among others). There are several reasons for its use. First, the logit model has been widely used and taught.[2] Second, it is relatively easy to understand and readily available in virtually all software packages. Finally, logit has resulted to be a fairly robust and reliable tool for forecasting financial distress.

We measure the incidence of bankruptcy with a binary random variable y whose realizations can be represented as

$$y = \begin{cases} 1 & \text{if } \theta' X \geqslant \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where

$$P(y = 1|x) = h(\theta' X). \tag{2}$$

The probability that the binary dependent variable equals one given the covariates is equal to a probability transformation of the single index $\theta' x$. In principle, both the parameters of the single index θ and the probability transformation function h need to be estimated. The logit model, however, is fully parametric in assuming a known form of h. In particular, h is a logistic cumulative distribution function

$$h(\theta' X) = \frac{e^{\theta' X}}{1 + e^{\theta' X}}$$

With this assumption the parameter vector θ can be estimated consistently and efficiently by maximizing

$$L = \sum_{i=1}^{N} [y_i \ln(P_i) + (1 - y_i) \ln(1 - P_i)], \tag{3}$$

where $P_i = e^{\theta' x_i}/(1 + e^{\theta' x_i})$ and $i = 1, \ldots, N$. Our logit models differ in the choice of the predictor matrix X. We consider four different approaches to choosing X. The first approach is the standard in applications of the logistic regression: the step-wise procedure with pre-selection of data. Other approaches involve estimating classification trees.

Logit models are not the only bankruptcy prediction models we consider. While our paper focuses on the use of classification trees in selection of bankruptcy predictors for standard parametric models like logit, it is also straightforward to use the classification tree for bankruptcy prediction. The two pioneering studies using classification and regression trees for bankruptcy prediction are those of Frydman, Altman, and Kao (1985), and Marais, Patell, and Wolfson (1984) who employed it to assess loan classifications. The first mentioned study compared CART to the classification precision of two discriminant models. Overall the classification-tree models were found to perform best. On the other hand, Marais et al. (1984) compared their recursive partitioning results against those of a multinomial probit model. Interestingly, they concluded that in estimating loan classifications there was very little to choose between the two procedures. More recent applications of decision tree models in bankruptcy or corporate financial distress prediction in general include Chen (2011), Li et al. (2010) and Min and Jeong (2009).

3. Variable selection

Selection of predictor variables is an important step in all bankruptcy prediction studies. To date no unified theory has been generally accepted. Most of the previous studies used a brute empirical approach of initial choice of variables (based also on expert knowledge) followed by a step-wise procedure to select the variables in the final logit or discriminant model. Such a procedure is not statistically rigorous. Different sequencing and/or initial ordering of variables need not result in a unique selection. As an attempt to overcome this deficiency some authors started using data mining techniques (Cho et al., 2010; Shirata, 1998). These are also better suited to capture potential nonlinearities in the relations between financial distress and predictor variables.

We propose a novel approach to using CART in selection of bankruptcy predictors and their subsequent use in prediction models. This approach is compared to more conventional methods of variable selection for the logit model. Both approaches are described in detail in the next two subsections.

3.1. CART selection approach

CART builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). Our reasons for selecting CART from the family of artificial intelligence methods are similar to those of Li et al. (2010). From a practical point of view, one important advantage of decision trees in bankruptcy prediction is the ability to generate easily understandable decision rules. This is feature is not shared by many artificial intelligence approaches.

The classic CART algorithm was popularized by Breiman, Friedmann, Olshen, and Stone (1984) (see also Ripley, 1996). CART model is a flexible method for specifying the conditional distribution of a variable y, given a vector of predictor values X. Such models use a binary tree to recursively partition the predictor space into subsets where the distribution of y is successively more homogeneous. The terminal nodes of the tree correspond to the distinct

[2] Among the first to apply logit to the problem of bankruptcy were Santomero and Vinso (1977) and Martin (1977) who employed it to examine failures in the US banking sector. Ohlson (1980) applied it more generally to 105 bankrupt and 2058 non-bankrupt firms. Notable applications that followed include Zmijewski (1984), and Wilson (1992). Accuracy of classification ranged from 76% in the work of Zmijewski (1984), where he employed probit and weighted exogenous sample likelihood models to investigate firms listed on the American and New York stock exchanges from 1972 to 1978, to 96% in the study by Pantalone and Platt (1987), where the authors use logit analysis to determine the causes of banks bankruptcy in the US after the deregulation.